INTERSPEECH 2009 BRIGHTON

Accounting for the Uncertainty of Speech Estimates in the Complex Domain for Minimum Mean Square Error Speech Enhancement

Ramón Fernandez Astudillo, Dorothea Kolossa, Reinhold Orglmeister

Chair of Electronics and Medical Signal Processing, Berlin Institute of Technology, Germany
ramon@astudillo.com, d.kolossa@ee.tu-berlin.de, Reinhold.Orglmeister@tu-berlin.de

10.21437/Interspeech.2009-371

Abstract

Uncertainty decoding and uncertainty propagation, or error propagation, techniques have emerged as a powerful tool to increase the accuracy of automatic speech recognition systems by employing an uncertain, or probabilistic, description of the speech features rather than the usual point estimate. In this paper we analyze the uncertainty generated in the complex Fourier domain when performing speech enhancement with the Wiener or Ephraim-Malah filters. We derive closed form solutions for the computation of the error of estimation and show that it provides a better insight into the origin of estimation uncertainty. We also show how the combination of such an error estimate with uncertainty propagation and uncertainty decoding or modified imputation yields superior recognition robustness when compared to conventional MMSE estimators with little increase in the computational cost.

Index Terms: Wiener Filter, MMSE-STSA, MMSE-LSA, Uncertainty Propagation, Uncertainty Decoding

1. Introduction

Minimum mean square error (MMSE) based speech enhancement methods obtain an estimation of the clean signal as the center of probabilistic density, or expectation, of the posterior distribution of the clean signal given the available noisy data. This posterior distribution is obtained by applying the Bayes theorem to a given model of noise and speech interaction, defined by a prior distribution of the clean signal and a likelihood function relating noisy and clean data. Since the resulting posterior distribution can be seen as a probabilistic description of the value of the clean signal given the available information, it provides an excellent framework for the application of uncertainty propagation, or error propagation, techniques for robust automatic speech recognition (ASR). Uncertainty propagation techniques transform a given probabilistic description of the clean signal in the domain of speech enhancement into the feature domain where speech recognition takes place. There, this information can be combined with the statistical parameters of the recognizer for improved robustness [1, 2, 3, 4]. This work concentrates on uncertainty models for the short-time Fourier transform (STFT) domain, which were initially developed to improve the recognition results after source separation [3] and have also been successfully employed for speech enhancement in non-stationary noisy environments [4]. In this paper, we derive analytical estimation methods for the uncertainty of estimation generated when applying three well known single channel MMSE speech enhancement methods: The Wiener filter and the two Ephraim-Malah filters. We show that the use of a complex domain uncertainty model rather than the usual consideration of the amplitude domain uncertainty model [2], allows a better modeling of the estimation error, which takes into consideration the natural asymmetry of the uncertainty distribution in the amplitude domain.

2. Uncertainty of Estimation in Single Channel MMSE Estimators

2.1. Posterior Distributions of the Gaussian Model

The model most often used for single microphone MMSE speech enhancement is the Gaussian model under the assumptions of the signal being non stationary, non ergodic, and the noise being additive [5, 6, 7, 8]. Under such a model, a noisy signal $y(t)$ containing a mixture of clean speech $x(t)$ and noise $d(t)$, given by

$$y(t) = x(t) + d(t), \tag{1}$$

is transformed into the time-frequency domain by employing the discrete version of the short-time Fourier transform (STFT). Each resulting complex valued Fourier coefficient X of the clean signal $x(t)$ is then estimated from the corresponding Fourier coefficient Y of the noisy signal $y(t)$ and the statistical parameters of the clean signal and noise Fourier coefficients X and D. The Gaussian assumption implies considering the real and imaginary components of X to be statistically independent, zero mean and Gaussian distributed with variance $\lambda_X/2$, which is equivalent to considering X as following a zero mean circularly symmetric complex Gaussian distribution of variance λ_X, given by

$$p(X) = \frac{1}{\pi\lambda_X} \exp\left(-\frac{|X|^2}{\lambda_X}\right). \tag{2}$$

The same model is also chosen for the noise signal D with variance λ_D. Due to the linearity of the STFT operator, the assumption of additivenes of noise in Eq. 1 implies that the likelihood function $p(Y|X)$, which relates noisy and clean data, will correspond to that of the complex Gaussian distribution of D, but centered on X

$$p(Y|X) = \frac{1}{\pi\lambda_D} \exp\left(-\frac{|Y-X|^2}{\lambda_D}\right). \tag{3}$$

From this statistical model, and by applying Bayes theorem, we can obtain various MMSE estimators like the Wiener and Ephraim-Malah filters. The Wiener filter corresponds to the solution obtained by individually solving the MMSE estimation of the real and imaginary components of X. Since the distributions in the model given by Eqs. 2 and 3 are complex Gaussian distributed, their real and imaginary components will be Gaussian distributed and the resulting posterior distribution of each

Copyright © 2009 ISCA

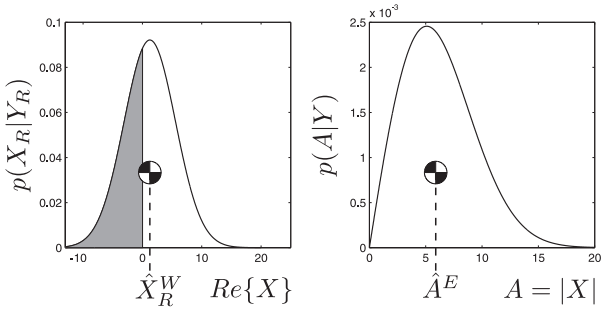

Figure 1: *Left: Posterior distribution (Eq. 4) and MMSE estimation $\hat{X}_R^W$ of the real component of a clean signal Fourier coefficient X (Wiener filter). Right: Posterior distribution (Eq. 7) and MMSE estimation $\hat{A}^E$ of the amplitude of a clean signal Fourier coefficient X (MMSE-STSA estimator). Shaded: Negative values of the real component posterior.*

component will be Gaussian as well. If we consider for example the real components of a Fourier coefficient of the clean and noisy signals X_R, Y_R, the MMSE estimation of the real component of the clean Fourier coefficient corresponds to the centroid of the posterior (see Fig. 1, left)

$$p(X_R|Y_R) = \frac{1}{\sqrt{\pi\lambda}} \exp\left(-\frac{\left(X_R - \frac{\lambda_X}{\lambda_X+\lambda_D} Y_R\right)^2}{\lambda} \right), \quad (4)$$

where the parameter λ was defined in [6, Eq. A.3] as

$$\frac{1}{\lambda} = \frac{1}{\lambda_X} + \frac{1}{\lambda_D} \quad (5)$$

Equivalently, the Wiener estimator for the imaginary part and its corresponding posterior are obtained by replacing X_R and Y_R by their imaginary counterparts X_I, Y_I. Then, the solution of the Wiener filter [5, Eq. 11] is obtained as

$$\hat{X}^W = \hat{X}_R^W + j\hat{X}_I^W = E\{X_R|Y_R\} + jE\{X_I|Y_I\} = \frac{\lambda_X}{\lambda_X + \lambda_D} Y, \quad (6)$$

and provides both an estimator for the amplitude of the clean signal and its phase.
Ephraim-Malah filters can be obtained by first noting that the phase of the clean signal plays a lesser role in the audible quality of the clean signal and therefore can be marginalized from the prior and likelihood function, given by Eqs. 2 and 3, which after applying Bayes theorem, leads to the following Rice distributed posterior (see Fig. 1, right)

$$p(A|Y) = \frac{2A}{\lambda} \exp\left(-\frac{A^2 + \left|\frac{\lambda_X}{\lambda_X+\lambda_D} Y\right|^2}{\lambda} \right) I_0\left(\frac{2|Y|A}{\lambda_D}\right), \quad (7)$$

where I_0 corresponds to the modified Bessel function of order zero. The first Ephraim-Malah filter corresponds to the centroid of this posterior [6, Eq. 7] and its therefore termed MMSE estimator of the short time spectral amplitude, or MMSE-STSA estimator. The second Ephraim-Malah filter corresponds to an MMSE estimation of the log spectral-amplitude and its therefore termed MMSE-LSA estimator. However, the posterior of the MMSE-LSA estimator, $p(\log A|Y)$, is never derived explicitly since the posterior expectation can be obtained through the moment generation function and the moments of Eq. 7 [7, Eq. 6]. To complete the estimation of the Fourier coefficient of the clean signal, the Ephraim-Malah amplitude estimation $\hat{A}$, is usually combined with the optimal MMSE estimator of the phase $\hat{\alpha}$, for the given model [6], which is equal to the phase of the noisy Fourier coefficient Y

$$\hat{X} = \hat{A}e^{j\hat{\alpha}} = \hat{A}\frac{Y}{|Y|}. \quad (8)$$

It is also worth noting that this phase estimator coincides with the phase estimator of the Wiener filter as given by Eq. 6.

2.2. Estimation Error and the Posterior Distribution

It is well known that the MMSE estimation error can be easily related to the variance of the posterior distribution. If we consider for example the MMSE-STSA estimator, we have that the estimated clean amplitude of the Fourier coefficient $\hat{A}^E$ is obtained as the center of probabilistic density of Eq. 7 (see Fig. 1, right)

$$\hat{A}^E = E\{A|Y\}. \quad (9)$$

Furthermore, the error of estimation can be expressed in terms of the prior and likelihood function as

$$E\left\{\left(A - \hat{A}^E\right)^2\right\} = \int\int \left(A - \hat{A}^E\right)^2 p(A|Y)p(Y)dAdY = \int \mathrm{Var}\{A|Y\}p(Y)dY, \quad (10)$$

where $\mathrm{Var}\{A|Y\}$ is the variance of the posterior distribution. When $\mathrm{Var}\{A|Y\}$ does not depend on the noisy data Y, the error of estimation corresponds directly to the variance of the posterior distribution. Unfortunately this is not the case for the Ephraim-Malah filters and the amplitude estimator obtained from the Wiener filter. Let us rather consider the error committed when estimating the clean signal in the complex domain. If we wish to perform a similar decomposition as in Eq. 10, but in the complex domain, we would first need to determine the posterior of the clean Fourier coefficient given the noisy Fourier coefficient $p(X|Y)$. In the case of the Wiener filter, this is trivial to obtain since real and imaginary components are independently estimated and their corresponding posteriors can be assumed to be statistically independent, Gaussian distributed and with same variance. Consequently, the posterior of the clean Fourier coefficient for the current model, obtained as

$$p(X|Y) = p(X_R|Y_R)p(X_I|Y_I) = \frac{1}{\pi\lambda}\exp\left(-\frac{\left|X-\hat{X}^W\right|^2}{\lambda}\right), \quad (11)$$

is circularly symmetric complex Gaussian distributed with mean equal to the Wiener filter estimation $\hat{X}^W$ and a variance equal to two times the variance of the posterior in Eq. 4. Since this variance is no longer dependent on Y, the expected error in complex domain corresponds to

$$E\left\{\left|X-\hat{X}^W\right|^2\right\} = \int\int\left|X-\hat{X}^W\right|^2 p(X|Y)p(Y)dXdY = \int \mathrm{Var}\{X|Y\}p(Y)dY = \int\lambda p(Y)dY = \lambda. \quad (12)$$

A more intuitive explanation for this is that the the non-linear dependency of the variance on the noisy signal Y in Eq. 7 is due to the amount of probabilistic density of $p(X|Y)$ reaching negative values (Fig. 1, left) which generates the asymmetry in the Rice posterior (Fig. 1, right)[1]. By considering the uncertainty in the complex STFT domain we get rid of this non-linear relation and thus are able to find a closed form solution for the estimation error.

Given this possibility it would also be desirable to find a similar closed form expression for the Fourier coefficient estimation error obtained when combining Eq. 8 with the MMSE-STSA and MMSE-LSA amplitude estimators. In order to do this we note that, on the one hand, the distance between Wiener and MMSE-STSA or MMSE-LSA solutions tends to zero for high SNRs and that this distance can be regarded as small compared to the variance of the amplitude posterior[2]. On the other hand, the phase estimator in Eq. 8 coincides with that of the Wiener filter. Regarding the MMSE-STSA estimator, we note also that the amplitude posterior in Eq. 7 corresponds to the distribution of the amplitude of the complex random variable X in Eq. 11. Taking this into account we can approximate the distribution $p(X|Y)$ for the MMSE-STSA Fourier coefficient estimation $\hat{X}^E$ by Eq. 11 but with mean equal to $\hat{X}^E$, rather than $\hat{X}^W$, leading to the following approximation of the estimation error

$$E\left\{\left|X-\hat{X}^E\right|^2\right\} \approx \lambda. \quad (13)$$

Regarding the MMSE-LSA estimator it is not possible to infer the form of the posterior distribution of the amplitude, although it can be assumed that it will be close to the Rice distribution since it is the result of propagating the posterior of the log-amplitude estimation $p(\log A|Y)$ into the amplitude domain by using the exponential. Subsequently the same approximation is also taken for the MMSE-LSA estimation error.

[1] For very high signal to noise ratios (SNR), the Rice distribution in Eq. 7 converges to a Gaussian distribution.

[2] This is also supported by the empirical determination of uncertainty distribution performed in [9].

3. Propagation of Uncertainty through the Feature Extraction

Given the result of previous section we can replace each estimated Fourier coefficient $\hat{X}$, obtained either with the Wiener or Ephraim-Malah filters, by the complex Gaussian distribution given by Eq. 11 with mean equal to $\hat{X}$, representing the uncertainty derived from the estimation process and thus obtaining a probabilistic description of the STFT of the clean signal. Methods for the propagation of first and second order moments of a probabilistic description of the short-time spectral amplitude (STSA) through the mel-cepstral feature extraction[3] are already well known and established [10, 3] and result in a probabilistic description for each cepstral feature given by its mean μ^{CEPS} and variance Σ^{CEPS}. Propagation of a complex STFT uncertainty model just requires an additional previous step to propagate the posterior distribution through the magnitude or magnitude squared transformations. A closed form solution for the propagation through the magnitude transformation using the moments of the Rice distribution was given in [4] and corresponds to

$$\mu^{\text{STSA}} = \frac{\sqrt{\pi\lambda}}{2}\cdot \mathrm{L}_2^1\left(-\frac{\hat{A}^2}{\lambda}\right), \quad (14)$$

$$\Sigma^{\text{STSA}} = \lambda + \hat{A}^2 - \left(\mu^{\text{STSA}}\right)^2, \quad (15)$$

where $\mathrm{L}_2^1(x)$ is the Laguerre polynomial that can be expressed in terms of modified Bessel functions. $\hat{A}$ corresponds to the estimated clean amplitude obtained either with Wiener, MMSE-STSA or MMSE-LSA estimators. The propagation through the magnitude squared transformation, to obtain the power spectral density (PSD), corresponds to [9]

$$\mu^{\text{PSD}} = \lambda + \hat{A}^2, \quad (16)$$

$$\Sigma^{\text{PSD}} = 2\lambda\hat{A}^2 + \lambda^2. \quad (17)$$

The computational complexity of the propagation methods can be considerably reduced if only diagonal covariances for the uncertainty are considered. In this case, the uncertainty propagation methods have computational needs comparable with twice that of conventional feature extractions.

4. Use of Uncertainty in Recognition Domain

Although many different methods exist that employ a probabilistic description of the features for robust ASR, we concentrate here on two well established alternatives used in the context of hidden Markov model (HMM) based ASR. Modified imputation estimates the clean observation o for the q^{th} Gaussian mixture of the recognizer, defined by its mean and variance μ_q and Σ_q, by maximizing the likelihood of the observation given mixture parameters and the uncertain description of the cepstral features, given by μ^{CEPS} and Σ^{CEPS} [3]. The corresponding estimator for each feature realizes

$$\hat{o}^q = \frac{\Sigma^{\text{CEPS}}}{\Sigma_q + \Sigma^{\text{CEPS}}}\mu_q + \frac{\Sigma_q}{\Sigma_q + \Sigma^{\text{CEPS}}}\mu^{\text{CEPS}}. \quad (18)$$

[3] A similar approach can also be applied for the Perceptual-Linear-Prediction features and RASTA filtered signals [9].

Another possibility is to directly compute the observation probability for the q^{th} Gaussian mixture by marginalizing over all possible unseen cepstra, thus obtaining the uncertainty decoding approach [1], which yields

$$\hat{p}(o|q) = N(\mu^{\text{CEPS}}|\mu_q, \Sigma_q + \Sigma^{\text{CEPS}}). \quad (19)$$

5. Experiments and Results

Table 1: *Word error rates (WER) for additive noise tests using white noise at different SNR values.*

SNR	∞	15	10	5	0
Wiener Filter	1.4	2.0	3.9	10.3	19.4
+Modified Imputation	**1.3**	**1.8**	**2.9**	**7.7**	**15.4**
+Uncertainty Decoding	**1.3**	1.9	3.0	8.4	16.7
MMSE-STSA	1.4	1.9	3.7	10.1	19.2
+Modified Imputation	**1.3**	**1.7**	**2.7**	**7.1**	**15.0**
+Uncertainty Decoding	1.4	**1.7**	2.8	8.0	16.4
MMSE-LSA	1.3	1.9	3.9	10.0	19.2
+Modified Imputation	1.3	**1.7**	**2.9**	**7.6**	**15.9**
+Uncertainty Decoding	1.3	**1.7**	3.0	8.3	17.2

Table 2: *Word error rates (WER) for additive noise tests using instationary street noise at different SNR values.*

SNR	∞	10	5	0	-5
Wiener Filter	1.4	2.7	5.9	11.9	21.0
+Modified Imputation	**1.3**	**2.1**	**3.8**	**9.6**	**18.0**
+Uncertainty Decoding	**1.3**	**2.1**	4.1	9.8	18.3
MMSE-STSA	1.4	2.4	4.8	11.3	21.3
+Modified Imputation	**1.3**	**2.1**	**3.6**	**8.6**	**17.7**
+Uncertainty Decoding	1.4	**2.1**	3.8	9.2	18.2
MMSE-LSA	1.3	2.4	5.2	11.1	20.9
+Modified Imputation	1.3	**2.1**	**3.7**	**9.0**	**17.8**
+Uncertainty Decoding	1.3	**2.1**	3.9	9.3	18.3

The efficiency of the proposed uncertainty estimation method was tested in automatic speech recognition tests using AURORA5 and JEIDA databases resources. Wiener, MMSE-STSA and MMSE-LSA estimators, modified to take into account speech presence probability [5, 8], were used in combination with the implementation of the mel-cepstral feature extraction and uncertainty propagation described in [4]. Only diagonal covariances where propagated. Improved minima controlled recursive averaging (IMCRA) was used for estimating noise power [11]. To provide a good reference baseline when only using the MMSE methods investigated[4], noises of the JEIDA database as well as white noise where used instead of AURORA5 noisy files. Noise files where artificially added to the clean AURORA5 files using segmental SNR as defined in [11, Eq. 35]. Training was carried out using a MATLAB version of the AURORA5 task train and test routines. Models were trained with the whole clean speech corpus using the AURORA5 scripts but the number of reestimations were limited to one after each mixture splitting. Tests were carried out with all the files of ten randomly picked speakers which proved to be representative of the total test corpus. Speech recognition was carried out by modifying the HTK toolbox to perform modified imputation and uncertainty decoding. The use of uncertainty propagation improved the recognition results in a majority of the tested scenarios, as shown in Table 1.

[4]The advanced front end [12], intended to be used with the AURORA5 database, combines multiple techniques for optimal robustness.

6. Conclusions and Discussion

We have introduced a new method for the computation of uncertainty of estimation of MMSE speech enhancement estimators based in the complex domain. In this domain the distribution of uncertainty can be regarded as symmetric and independent of the available noisy Fourier coefficient, allowing us to compute the exact MMSE estimation error in the case of the Wiener filter and to approximate it for the MMSE-STSA and MMSE-LSA estimators. We have also demonstrated that such an approach in combination with uncertainty propagation techniques improves the robustness of ASR systems. It should be noted, however, that the error measure considered does not account for the errors committed when estimating the noise variance λ_D, and the consequent error in computing λ_X. A more general framework which includes these errors along with other factors, like speech presence uncertainty, is already under investigation. Under such a model, a higher variance of the amplitude estimation for low SNRs can be expected, further substantiating the assumptions made in Sec. 2.2. Finally, a more generalized model is also expected to improve the recognition results furthermore.

7. References

[1] Droppo, J., Acero, A. and Deng, L. "Uncertainty decoding with SPLICE for noise robust speech recognition", Proc. IEEE ICASSP02, 1, 57–60, 2002

[2] Stouten, V. and Van hamme, H. and Wambacq, W., "Model based feature enhancement with uncertainty decoding for noise robust ASR", Speech Communication., 48(11), 1502–1514, 2006

[3] Kolossa, D. and Sawada, H. and Astudillo, R.F. and Orglmeister, R. and Makino, S., "Recognition of Convolutive Speech Mixtures by Missing Feature Techniques for ICA", Proc. Asilomar06, 1397–1401, 2006

[4] Astudillo, R.F., Kolossa, D. and Orglmeister, R., "Propagation of Statistical Information through non-linear Feature Extractions for Robust Speech Recognition", Proc. MaxEnt07, 245–252, 2007

[5] McAulay, R. J., Malpass, L. M., "Speech enhancement Using a Soft-Decision Noise Suppression Filter",Proc. IEEE Trans. Speech and Audio, 28(2): 137–145, 1980

[6] Ephraim, Y., Malah, D., "Speech enhancement using a minimum mean-square error short-time amplitude estimator", Proc. IEEE Trans. Speech and Audio, 32(6), 1109–1121, 1984

[7] Ephraim, Y., Malah, D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", Proc. IEEE Trans. Speech and Audio, 33, 443-445, 1985

[8] Cohen I., Berdugo, B., "Speech enhancement for non-stationary noise environments", Proc. IEEE Trans. Speech and Audio, 11(5), 466–475, 2003

[9] Astudillo, R.F., Kolossa, D. and Orglmeister, R., "Uncertainty Propagation for Speech Recognition using RASTA Features in Highly Nonstationary Noisy Environments", Proc. ITG-Fachtagung, 61, 2008

[10] M. J. F. Gales., "Model-Based technique for noise robust speech recognition", P.h.D. thesis, Gonville and Caius College, 1995.

[11] Cohen I., "Noise Spectrum Estimation in Adverse Environments: Improved Minima controlled Recursive Averaging", Proc. IEEE Trans. Speech and Audio, 11(5), 466–475, 2003

[12] "Advanced front-end feature extraction algorithm; ES 202 050 V1.1.1", European Telecommunications Standards Institute, 2002